Continuous Speech Recognition System for Kannada Language with Triphone Modelling using HTK

Anand H. Unnibhavi, D.S.Jangamshetti, Shridhar K.

Abstract*: *Abstract: Kannada is the regional language of India spoken in Karnataka. This paper presents development of continuous kannada speech recognition system using monophone modelling and triphone modelling using HTK. Mel Frequency Cepstral Coefficient (MFCC) is used as feature extractor, exploits cepstral and perceptual frequency scale leads good recognition accuracy. Hidden Markov Model is used as classifier. In this paper Gaussian mixture splitting is done that captures the variations of the phones. The paper presents performance of continuous Kannada Automatic Speech Recognition (ASR) system with respect to 2, 4,8,16 and 32 Gaussian mixtures with monophone and context dependent tri-phone modelling. The experimental result shows that good recognition accuracy is achieved for context dependent tri-phone modelling than monophone modelling as the number Gaussian mixture is increased.

Keywords* : *About four key words or phrases in alphabetical order, separated by commas.

I. INTRODUCTION

In recent years, ASR (Automatic Speech Recognition) techniques have taken a great leap forward with the help of Deep Neural Network (DNN) based approaches Signal modelling and pattern matching [1] are the basic operation of ASR. Accuracy in ASR is a major challenge because of change in context, speakers and noise in the environment. Two main components of ASR are, feature extractor and classifiers.

There are many Classifiers used for building ASR such as Hidden Markov Model (HMMs), neural networks and deep neural networks [2]. HMMs are most widely used in speech recognition. Speech can be thought of as a Markov model for many stochastic purposes. Another reason of popularity of HMMs is that they can be trained automatically. HTK, Julius, Sphinx, and Kaldi [3] are the ASR tools. There is a lot of scope to develop automatic continuous speech recognition system for Kannada language, which is one of the oldest languages of our country. Historically languages like marathi, telagu, tamil etc. have been derived from Kannada language. In this work MFCC is used as feature extractor and HTK as ASR tool. The paper consists of different sections and as follows: Section 2 deals with Literature survey. Section 3 deals with proposed method, implementation, and results section 4 deals with conclusions.

Revised Manuscript Received on September 15, 2019

Anand H. Unnibhavi, Dept. of Electronics and Communication Basaveshwara Engineering College ,Bagalkot-587103, India
Email: anandhu.rampur@gmail.com
D.S.Jangamshetti , Dept. of Electrical and Electronics Basaveshwara Engineering College, Bagalkot-587103, India
Email: asdj1229@gmail.com
Shridhar K. Dept. of Electronics and Communication Basaveshwara Engineering College, Bagalkot-587103, India
Email: shridhar.ece@gmail.com

II. LITERATURE SURVEY

Gaurav *et al.* [4] presented technique to build a speaker independent continuous ASR for Hindi language. MFCC is used as speech feature parameter and HMM as classifier. HTK-3.4 is used as ASR tool. Ahmad *et al*. [5] designed and implemented English digits speech recognition system using Matlab. Two modules were developed, namely the isolated words ASR and the continuous ASR systems. It is compulsory to send paper in both email address. K. Sreenivasa Rao [6] demonstrated prosody models to build speech systems in Indian languages. Duration and intonation models developed using feed forward neural networks are considered as prosody models. In this study author has considered, 145 CV units (classes) for the recognition. Support vector machines (SVMs) are used for the recognition of CV units. MFCC is used as feature vector.

N. Uma Maheswari *et al.* [7] described speaker independent ASR system for Indian English language. Phoneme recognition is built using neural network. The second module built using HMM based on timit corpus. Muralikrishna H. *et al.* [8] discusses the implementation of Kannada isolated digit ASR that uses MFCC as feature vector and Hidden Markov Model (HMM) as pattern recognizer. First and second order derivatives are used to obtain better accuracy.

M.A. Anusuya and S.K. Katti [9] developed a statistical method to eliminate the silence from the audio signal, based on vector quantization method. MFCC is used to extract the features. Recognition error rate reduced from 2.59 to 1.56 by using VQ1 clustering algorithm and 2.5 to 1.45 by using VQ2 algorithm. Also the speaker dependent error recognition rate has also been decreased from 2.5 to 1.45, S.B. Harisha *et al.* [10] presents a Kannada speech recognition system. MFCC is used to extract the features of segmented words along with KNN as classifier. Out of 10 sentences 7 sentences are recognized correctly and out of 104 words 95 words are recognized correctly. For sentences (connected words) recognition accuracy is 91.5%. Thimmaraja Yadava G. and H. S. Jayanna [11] have proposed speech recognition work using Kaldi which involves MFCC as feature extractor.

The speech data is collected from 1100 farmers across Karnataka state. Speech data is recorded from 850 males and 250 females. It includes 180 isolated words. Language Models (LMs) and Acoustic Models (AMs) are developed by using 10000 district name utterances of Karnataka districts. 16000 mandi name utterances are used for Karnataka mandis. 18100 commodity name utterances are used. The Word Error Rate (WER) using Kaldi for Kannada language are 9.34%, 11.10%, 11.40% and 11.25% for districts, mandis, commodities and overall speech data respectively.

Through this survey it is found that some of the major challenges in ASR include better recognition rate, low word error rate, and developing speech database for regional language. It is also found that classifiers such as HMM, Support Vector Machine (SVM), DNN and KNN yield good accuracy for isolated speech recognition system and Continuous speech recognition system which is around 90% irrespective of the language. Hence there is a need to develop a new framework for Kannada language recognition which combines the advantages of MFCC, HMM and HTK ASR as tool.

III. PROPOSED METHOD

This section deals with the details of proposed work implementation.

A. HTK Processing Stages

ASR system is implemented using HTK 3.4.1 version. Speech corpus is needed both for training and testing. In the case of training data, the transcriptions of the training utterances are used in conjunction with pronunciation dictionary. Here Kannada acoustic-phonetic database is used for implementation. ASR using HTK involves four basic steps as shown in Fig. 1

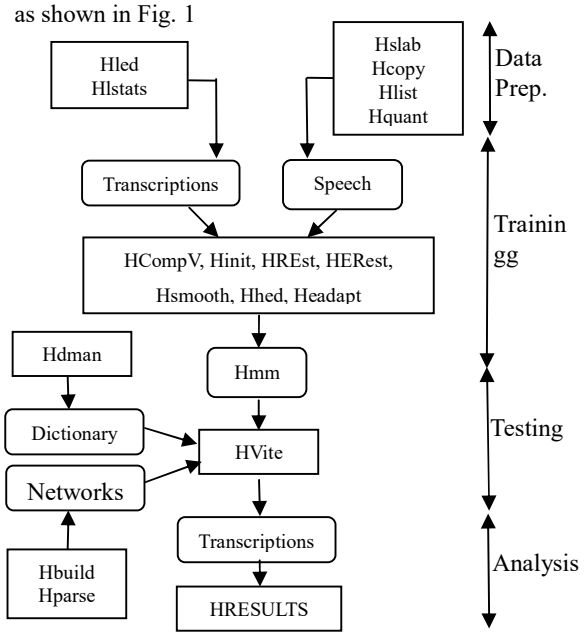

Fig. 1: HTK Processing stages for ASR

B. Data Preparation

In this work, four native female speakers are choosen for sentence utterance. Each speaker is asked to speak 300 Kannada sentences. A total of 1200 (4 speakers x 300 sentences = 1200 sentences) sentences are recorded with the help of wave surfer tool. The spoken sentences are sampled at 16 kHz with 16 bit depth. The word network is created by grammar file using the **HParse** tool, Master Label File (MLF) is a single file containing complete set of transcriptions word label and phone label are created by editor **HLEd** [12]. The step called acoustical analysis is the final stage of data preparation. Mel Frequency Cepstral Coefficients (MFCCs) features are derived using the tool **Hcopy.** 39 MFCC vectors are created and consists of 13 MFCC coefficients, 13 delta coefficients and 13 delta-delta coefficients.

C. Training Phase

The first step in HMM training is to define a prototype model. Its purpose is to define the model topology. A total of 1180 sentences are used for training purpose and resulting in 64 distinct phones generated from the collected sentences, so a total of 64 HMMs are created. For phone-based systems, a good topology to use 3 active states and two non emitting states. The **HInit** is the HTK tool that uses the predefined prototype, acoustic vector. The HTK tool **HCompV** will compute the global mean and variance and set all of the Gaussians in a given HMM to have the same mean and variance. The flat start monophones stored in the directory hmm0 are re-estimated using the **HERest,** this tool updates to maximize the probability of observation sequence. This retraining is performed for several iterations.

Phones are context dependent based on the preceding and succeeding phones. Monophone modeling cannot capture the phone variation. This shortcoming of monophone is overcome by triphone modelling and yields better performance of the ASR.

D. Tied-State Triphone

Monophone-based acoustic models never capture the context dependent phones that give poor recognition accuracy, leading to the work of triphone acoustic modelling. For a given set of monophone HMMs, the final stage of model implements context dependent triphone HMMs. Triphone creation is as follows. In the first case, triphone transcriptions are derived from monophone transcription, triphone models are generated by copying the monophones and reestimating. In the second case, acoustic states for the triphones created are tied together. Database of 1200 sentences with 2264 words and 3114 triphones are created and these are called logical HMM's. Good approximation is achieved when the parameters are reestimated, which results in better recognition accuracy then monophone modelling.

E. Gaussian Mixture Model

Gaussian Mixture Model (GMM) provides a statistical representation of how speaker generate sounds.

Single Gaussian model do not capture the distribution of feature vectors in a better way, So Gaussian mixture splitting is performed. These multi-Gaussian models will capture all the variations of phone [13]. The general form of a mixture model is

$$p(x/\theta) = \sum_{m=1}^{M} p(x, w_m / \theta_m) = \sum_{m=1}^{M} c_m p(x/w_m, \theta_m) \quad (1)$$

'x' is the observed variable' θ 'is the hidden variable. A d-dimensional (GMM) with M components, μ_m mean vectors c_m mixture weights, and Σ_m of covariance matrices. The density function of a Gaussian mixture model is.

$$p(x/\theta) = \sum_{m=1}^{M} c_m N(x; \mu_m, \Sigma_m) \quad (2)$$

where c_m is the component prior of each Gaussian component. For this to be a valid probability density function it is necessary that

$$\sum_{m=1}^{M} c_m = 1 \quad \text{and} \quad c_m \geq 0 \quad (3)$$

Increasing the number of components will get large number of parameters and results in better Probability Density Function (PDF) and results in better classifier. For maximum likelihood estimation, 'n' need to be specified.

$$l(\theta) = \sum_{i=1}^{n} \log p(x_i/\theta) = \sum_{i=1}^{n} \log \left[\sum_{m=1}^{M} c_m p(x_i/w_m, \theta_m) \right] \quad (4)$$

w_m is the mixture weight. Therefore d-dimensional continuous-valued data vector (features) is

$$l(\theta) = \sum_{i=1}^{n} \log \left[\sum_{m=1}^{M} \frac{c_m}{(2\pi\sigma_m^2)^{d/2}} \exp\left(\frac{-\| x_i - \mu_m \|^2}{2\sigma_m^2} \right) \right] \quad (5)$$

$$\mu_m = \frac{\sum_{i=1}^{n} p(w_m/x_i, \theta) x_i}{\sum_{i=1}^{n} p(w_m/x_i, \theta)} \quad (6)$$

$$\sigma_m^2 = \frac{1}{d} \frac{\sum_{i=1}^{n} p(w_m/x_i, \theta) \| x_i - \mu_m \|^2}{\sum_{i=1}^{n} p(w_m/x_i, \theta)} \quad (7)$$

It is essential to maximize the expectation in order to maximize the likelihood function. Expectation is maximized as follows and it is an iterative process. Let the parameters at iteration k be denoted by θ^k

$$\mu_m^{k+1} = \frac{\sum_{i=1}^{n} p(w_m/x_i, \theta^k) x_i}{\sum_{i=1}^{n} p(w_m/x_i, \theta^k)} \quad (8)$$

$$\sigma_m^{2(k+1)} = \frac{1}{d} \frac{\sum_{i=1}^{n} p(w_m/x_i, \theta^k) \| x_i - \mu_m^{k+1} \|^2}{\sum_{i=1}^{n} p(w_m/x_i, \theta^k)} \quad (9)$$

$$c_m^{k+1} = \frac{1}{n} \sum_{i=1}^{n} p(w_m/x_i, \theta^k) \quad (10)$$

One of the uses of above equations is to maximize the likelihood function thus

$$l(\theta^{(k+1)}) \geq l(\theta^{(k)}) \quad (11)$$

Therefore the overall procedure is

i. Initialize the parameters of the mixture, θ^0 for examle set all component priors to be equal, all variances to be equal, and use different values for the mean vectors and set k = 0

ii. Calculate $\sum_{i=1}^{n} p(w_m/x_i, \theta^k)$ for all data point and collect the statistics for the numerators and denominators of the re-estimation formulae. Also calculate the log likelihood of the data set

iii. Update the parameters as necessary to give θ^{k+1}, set k = k + 1

iv. If the log likelihood increase is less than a threshold, stop, else go to step 2. Further iterations will increase the likelihood further.

F. Testing Phase

In this stage a new test transcription will be generated for an unknown 20 test sentences as it is done in training corpus preparation. The test signal will be converted in to a series of acoustic vector that is .mfcc files using HTK tool **Hcopy**. Another HTK recognition tool **HVite** is used to generate an output transcription file in master label file (.mlf) format by taking the input observation that is acoustic vectors along with HMMs definition, Kannada word dictionary, and HMM list. Also the above command when executed fetches stored HMMs found in hmm1 folder to convert input word level transcription (words.mlf) with new phone level transcription (aligned.mlf). The **HVite** command with Viterbi algorithm will match with the recognizer's Markov models. Words.mlf file is again processed using filtering module that elicits the recognized word, sentences and exhibit in text form. The recognizer is complete and its performance can be estimated.

G. Results

Total of 20 test utterances files are used to compute the performance of ASR, features of these 20 files are extracted using **HCopy** tool with same configuration parameters that are used while training and stored as test.scp files, then each test file is recognized along with its transcription output to the recout.mlf, test.mlf contains word level transcriptions for each input test file, the performance is estimated by using **HResults** tools**.** The tool gives word, sentence level statistics indicating recognition accuracy.

$$\text{Percent of word Correct} = \frac{N - D - S}{N} \times 100\% \quad (12)$$

$$\text{Percent of sentence Accuracy} = \frac{N - D - S - 1}{N} \times 100\% \quad (13)$$

Total number of test spoken sentences is N, number of deletions is D, number of insertions is I, S stands for substitutions. Analysis of the ASR is conducted with five iterations of varying Gaussian mixtures (2, 4, 8, 16, and 32). Figure 2 shows Recognition accuracy for sentence level monophone and triphone.

From the figure it is observed that recognition accuracy varies from 59.59% to 85.14% and for context-dependent triphone recognition accuracy varies from 73.97% to 95.17% as the number of Gaussian mixtures varies 2 to 32. Therefore triphone model gives better results as compared to the monophone model.Similarly for word level analysis, experiments were performed five times with different number of Gaussian mixtures. Figure 3 shows recognition accuracy for context-independent (monophone) and context dependent (triphone) models. It is observed that recognition accuracy varies from 73.29% to 82.24% and 83.56% to 91.78% for monophone and triphone respectively. Better word level accuracy is obtained for context based-triphone model when compared with monophone modelling. Figures 4 and 5 show relative improvement in recognition accuracy for sentence level and word level respectively for monophone and triphone modelling. From the figure it is observed that not much increment in recognition accuracy, as the number of Gaussian mixtures are increased. It is mainly due to over fitting of Gaussian mixture model parameters.

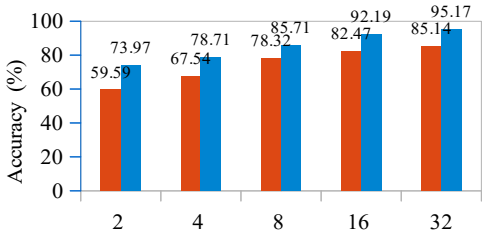

Fig. 2 : Monophone and Triphone sentence recognition accuracy

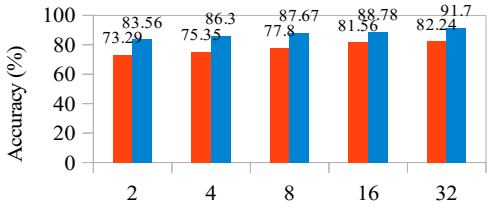

Fig. 3 : Monophone and Triphone word level recognition accuracy

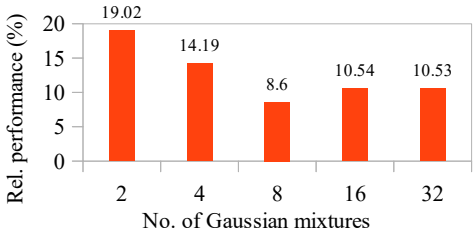

Fig. 4 : Relative Performance for sentences

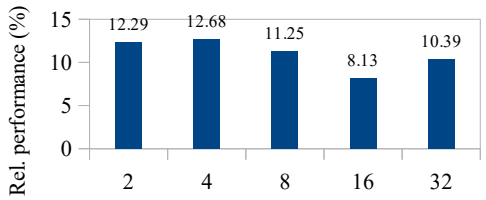

Fig. 5 : Relative Performance for words

IV. CONCLUSION

Kannada Continuous ASR system using triphone modelling developed is implemented using MFCC and HTK with monophone modelling and triphone modelling in combination with Gaussian mixture. Sufficient amount of speech Corpus is developed for training and testing purpose. Single Gaussian model doesn't capture all the variation of the phones. So Gaussian mixture splitting is performed. These multi-Gaussian models will capture all the variations of phone. Testing is performed using Viterbi search process and from experimental results shows that tied-state triphone system gives better performance then monophone system and accuracy increases as the number of mixture components is increased. In this work context-independent results in 85.14% accuracy and context-dependent yields 95.17%.

REFERENCES

1. Anand H. Unnibhavi and D.S. Jangamshetti, "A Survey of Speech Recognition on South Indian Languages", International conference on Signal Processing, Communication, Power and Embedded System IEEE (SCOPES)-2016.W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
2. Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition",IJFARIEAT, Vol.1(6), July 2013.
3. Hans Krupakar and Keerthika Rajvel, "A survey of voice translation methodologies - acoustic dialect decoder", International Conference On Information Communication and Embedded Systems (ICICES-2016).
4. Gaurav, Devanesamoni Shakina Deiv, Gopal Krishna Sharma and Mahua Bhattacharya, "Development of Application Specific Continuous Speech Recognition System in Hindi", Journal of Signal and Information Processing, 2012, 394-401, Published Online August 2012. J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.
5. Ahmad A. M. Abushariah, Teddy S. Gunawan and Othman O. Khalifa, "English Digits Speech Recognition System Based on Hidden Markov Models", International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia.
6. K. Sreenivasa Rao, "Application of prosody models for developing speech systems in Indian languages," Int J Speech Technol (2011) 14: 19–33 DOI 10.1007/s10772-010-9086-9.
7. N. Uma Maheswari, A. P. Kabilan and R. Venkatesh, "Speaker independent speech recognition system based on phoneme identification ,"Proceedings ofthe 2008 International Conference onComputing, Communication andNetworking (ICCCN 2008) 978-1-4244-3595-1/08.
8. Muralikrishna H. Ananthakrishna and T. Kumara shama, "HMM Based Isolated Kannada Digit Recognition System using MFCC ", 978-1-4673-6217-7/13/$31.00 2013 IEEE.

9. M.A. Anusuya and S.K. Katti, "Speaker Independent Kannada Speech Recognition using Vector quantization", MPGI National Multi Conference 2012 (MPGINMC-2012) "Advancement in Electronics & Telecommunication Engineering" 7-8 April, 2012 Proceedings published by International Journal of Computer Applications (IJCA)ISSN: 0975 - 8887 32.
10. S. B. Harisha, S. Amarappa and S. V. Sathyanarayana, "Kannada Speech Recognition Using MFCC and KNN Classifier for Banking Applications", International Journal of Innovative Research in Computer and Communication Engineering, Vol.5 (1), January 2017 Copyright.
11. Thimmaraja Yadava G. and H. S. Jayanna "Creating Language and Acoustic Models using Kaldi to Build An Automatic Speech Recognition System for Kannada Language 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology.
12. S. Young, "The HTK book (for HTK version 3.4)" Cambridge University Engineering Department, March 2009.
13. Ankit Kumar, Mohit Dua and Tripti Choudhary, "Continuous Hindi Speech Recognition Using Gaussian Mixture HMM", 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science.

AUTHORS PROFILE

Anand H. Unnibhavi completed his B.E in Electronics and communication Engineering from Vidya Vardhaka college of Engineering Mysore, affiliated to VTU Belagavi Karnataka and obtained his M.Tech in the area of Digital Electronics and Communcation system from Malnad College of Engineering Hassan , affiliated to VTU Belagavi Karnataka. Currently he is pursuing Ph.D in the area of Speech Processing. His Areas of interest are Speech processing, Wireless network. Presently working as Assistant professor in the department of Electronics and Communication Engineering, Basaveshwara Engineering College Bagalkot,Karnataka, India.

Dr. Dakshayani S. Jangamshetti was born in Ilkal, Karnataka, India on 12th February 1964. She obtained her B.E (Electrical) degree from Karnataka University Dharwad in 1985 and M.Tech.(Instrumentation) Degree from IIT Kharagpur in 1989 & Ph.D (Speech Processing) from IIT, Mumbai in 2003. Her areas of interest include speech signal processing, Image processing, Microcontroller and signal systems. She won the "Outstanding IEEE Branch Counselor" award for the year 2014. Presently, she is a Professor of Electrical and Elecronics Engineering department at Basaveshwar Engineering College (Autonomous), Bagalkot.

Dr. Shridhar S. Kuntoji received his B.E(Electronics and Communication Engineering) degree from Gulbarga University Gulbarga, M.Tech (Digital Electronics and advanced communication) from NITK Surathkal, and Ph.D in the area of speech signal processing from Shivaji University Kolhapur in the year 1988, 2000 and 2014 respectively. He joined as Lecture in Electronics and Communication Engineering, Basaveshwara Engineering College Bagalkot, India in the year 1993, where he is currently working as Professor since 2014. He is currently involved in the research area of speech enhancement focusing on people suffering from hearing loss.